VOCAL TRACT SETTINGS IN SPEAKERS WITH OBSTRUCTIVE SLEEP APNEA SYNDROME

J. L. Blanco [1], J. Schoentgen [2]

[1] Signal Processing Applications Group, Universidad Politécnica de Madrid
ETSI de Telecomunicación, Avda. Complutense 30, 28040 Madrid, Spain
jlblanco@gaps.ssr.upm.es

[2] National Fund for Scientific Research, Belgium & Laboratories of Images, Signal Processing and Acoustics
CP 165/51, Faculty of Applied Sciences, Université Libre de Bruxelles, 50, Av. F.D. Roosevelt, B-1050, Brussels, Belgium
jschoent@ulb.ac.be

Abstract: **Automatic systems based on speech signal analysis for the early detection of obstructive sleep apnea (OSA) have achieved fairly high performance rates in recent years. However, a satisfactory explanation of these results has not been available. This presentation aims at explaining via an examination of the long-term spectra of OSA patients and normal control speakers these systems' ability to discover OSA speakers on the base of all-purpose cepstral coefficients. An interpretation of the long-term spectra in terms of the underlying tract settings suggests that the speech of OSA patients is characterized by a pharyngeal narrowing that may be captured by acoustic cues of the spectral contour of windowed speech frames. A novel interpretation of long-term spectra in terms of the first principal component of the temporal sequence of short-term amplitude-spectra is also discussed.**

Keywords: obstructive sleep apnea (OSA), spectral analysis of speech, vocal tract settings, long-term average spectrum (LTAS), pharyngeal narrowing.

I. INTRODUCTION

Severe obstructive sleep apnea is characterized by the interruption of breathing during sleep [1], involving episodes that may last for more than 10 seconds and which recurrently occur during the night, with up to more than 30 episodes per hour. This syndrome affects 2 to 4% of the male population between 30 and 60 years of age. A possible effect of OSA is daytime sleepiness, increasing the risk of the patient to get involved in traffic accidents or leading to poor work performance [2].

The discovery via speech analysis of patients that have a propensity to suffer from severe obstructive sleep apnea syndrome (OSA) has become more reliable in recent years. Previous work by the first author has shown that it is possible to discriminate between modal speakers and speakers suffering from severe OSA by means of generic automatic classifiers that rely on a conventional coding of speech frames by means of mel-frequency cepstral coefficients ([3]–[5]). However, no results on speech analysis are available that would enable linking that observation to OSA speaker anatomy, physiology or OSA pathogenesis.

This presentation aims at explaining the ability to discover OSA patients via speech analysis in terms of what is known about speech production in general and vocal tract settings in particular. J. Laver [6] has discussed vocal tract settings extensively. They designate articulatory biases of the neutral tract shape, which are common to all the phonetic segments of the utterances of a speaker. The susceptibility of individual phonetic segments to tract settings is variable depending on the degree by which the properties of a phonetic segment are affected. For instance, one expects +spread sounds to be biased by a lip protrusion setting more than +round sounds that are characterized by lip protrusion anyway.

Easily observable anatomical or physiological OSA cues are patients' weight, height, body mass index and cervical perimeter as well as snoring. The latter together with other symptoms (occasional hypoplasia and/or backward displacement of the maxilla and mandible) suggest that their oro-pharyngeal cavity is narrowed. Magnetic Resonance Imaging of OSA patients has confirmed this and that sole observation can be used to identify OSA cases. The narrowing of the pharynx may be assimilated to a vocal tract setting that is likely to bias a speaker's speech sounds via a shift in vocal tract resonances. Experiments are reported hereafter that have been carried out with a view to testing that hypothesis on speech records.

J. Laver [6] and F. Nolan [7] have suggested tracking vocal tract settings in speech by means of the Long Term Average Spectrum (LTAS). One obtains the LTAS of an utterance by computing the amplitude spectrum for successive frames and averaging the amplitude spectra. Averaging is expected to lessen the contribution of phonetic segment variability and boost what is common to all frames (i.e. the tract setting or phonatory setting, depending on the frame length). In the past, Long Term Average Spectra often have been used to characterize a speaker overall, including information from all the speech frames. Focussing by means of the LTAS on tract settings only has been rare.

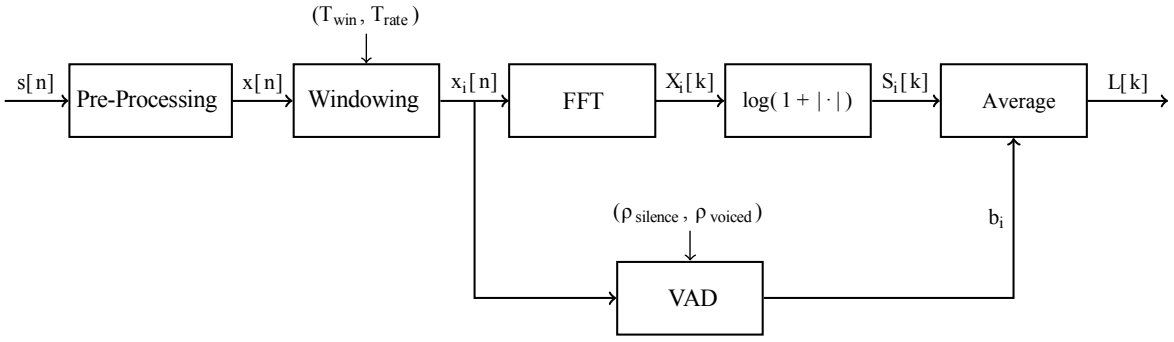

Figure 1: Block diagram of the long-term average spectra estimation.

II. METHODS

Our corpus has comprised speech records from 80 male speakers [8]. Forty of them have been healthy or have suffered from mild OSA (AHI<10), while the remaining forty have been patients suffering from severe OSA (AHI>30). The apnea-hypo-apnea index (AHI) is the number of apnea plus hypoapnea events per hour of sleep, and is often used by clinicians to describe the severity of patients' condition and to adjust their treatment [9]. Hereafter, the AHI<10 speakers are designated as controls and the AHI>30 speakers as patients. The two groups have been balanced for weight, height and age to decrease the influence of secondary speaker characteristics.

Each speaker sustained a complete set of Spanish vowels [i,e,a,o,u] as well as four Spanish sentences. The sentences have been short and phonetically balanced. They have been designed with a fixed number of intonation groups to decrease intra-speaker variability.

The computation of the LTAS has involved (i) pre-emphasizing the speech signal (α=0.99) to remove zero-frequency and ultra-low frequency spectral components, (ii) windowing, (iii) voicing detection, (iv) computing the amplitude spectrum, (v) boosting high-frequency amplitudes and (vi) averaging. The block diagram in Fig. 1 summarizes steps (i) to (vi).

The frame length has been set to T_{win}= 5 ms and the frame hop to T_{rate}= 3 ms to decrease the influence of the voice source harmonics on the amplitude spectrum because the focus is on the spectral contour that reports vocal tract resonances.

Even though other publications have retained vowel nuclei as the sole contributors to LTAS [10], here any voiced frames have been included. Voiced frames have been discovered by means of a voicing detector proposed in [11] and which is easily tunable by weighting each of its two steps that are the following. The first involves a frame-by-frame energy estimation. The purpose is to detect and remove silent intervals. Threshold $\rho_{silence}$ for speech activity detection has been set to 20% of the average energy of the frames of one speech record. The second step involves auto-correlation coefficient ρ_{ss} between an analysis frame and the analysis frame delayed by one sample. The purpose of that step is to detect the frames that are voiced and for which the auto-correlation coefficient is large, i.e. $\rho_{ss}(1) / \rho_{ss}(0) \geq 0.9$. Only frames that have been tagged as voiced have been used to compute the LTAS.

The purpose of boosting is to increase the amplitude of high-frequency components compared to low-frequency components. The reason is that the spectral slope of the voice source causes higher formants to be feeble and barely visible in the amplitude spectrum. The boosting function is the logarithm of the spectral amplitude+1. The +1 guarantees that all amplitudes have positive log-transforms that respect monotony so that the average of the boosted amplitudes can be interpreted meaningfully.

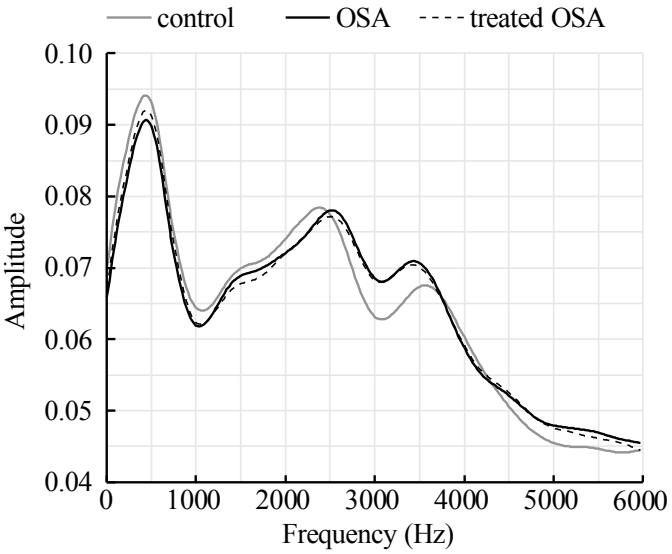

Figure 2: Average LTA spectra for control speakers (continuous, gray line), OSA patients (continuous, black line) and treated OSA patients (dashed, black line) of one Spanish sentence.

Table 1: Sign of the change of the frequencies of formants F1 to F4 when the cross-sectional area of the tube is slightly decreased with regard to the neutral tract shape. The lip and glottal regions are to the left and right respectively.

Upper Airway Regions	bilabial labiodental		dental / alveolar	post-alveolar palatal		velar		oropharyngeal		hypopharyngeal		epiglotto-pharyngeal	epilaryngeal	glottal
Formants	A	B	C	D	**E**	F	G	$\overline{\mathbf{G}}$	$\overline{\text{F}}$	$\overline{\text{E}}$	$\overline{\text{D}}$	$\overline{\text{C}}$	$\overline{\mathbf{B}}$	$\overline{\text{A}}$
F1	−	−	−	−	−	−	−	+	+	+	+	+	+	+
F2	−	−	−	+	+	+	+	−	−	−	−	+	+	+
F3	−	−	+	+	+	−	−	+	+	−	−	−	+	+
F4	−	+	+	+	−	−	+	−	+	+	−	−	−	+

III. Results

Fig. 2 shows the averages of the LTAS obtained for the control speakers (continuous, gray line), the OSA patients (continuous, black line) and treated OSA patients (dashed, black line). (Palliative) treatment consists in providing continuous positive air pressure during sleep to prevent airway collapse. Treatment does not modify significantly the configurations of the upper airway found in these patients (ca. 1.6 mm increase on average (i.e. 12%) according to [12]).

The average LTAS that are reported have been obtained for one sentence out of four. Results for the other three are similar. The results for sustained vowels are category-dependent and more difficult to interpret.

Fig. 2 evidences differences between OSA and control speakers with regard to the positions of the third and fourth formants, the distance between which is smaller for OSA than for control speakers. Formants F3 and F4 have respectively shifted up and down for the OSA patients. Palliative treatment does not appear to have an influence on the OSA speakers' vocal tract settings, in accordance with its small impact on the anatomy of OSA patients' vocal tracts.

IV. Discussion and Conclusion

The observed differences between OSA and control speakers are best explained in terms of the sensitivity functions of the vocal tract in the vicinity of the neutral vocal tract shape [13]. Table 1 shows the sign of the change of the frequencies of formants F1 to F4 when the area is slightly decreased with regard to the neutral tract shape. Regions labeled A to A-bar within which the signs of formant-specific sensitivity functions are the same actually have unequal lengths. These length differences are not reported in Table 1. One observes three regions the narrowing of which is susceptible to shift formants in agreement with what is observed in Fig. 2.

In Table 1, region B-bar corresponds to the epilarynx the narrowing of which has been linked to the singer's formant. Region G-bar corresponds roughly to the oro-pharynx the narrowing of which is expected for OSA patients. Region E agrees with the palate, to which no role is assigned within the framework of the present study.

One may therefore conclude that the LTAS enables tracking vocal tract settings and that the tract settings and therefore the timbre of OSA speakers tend to differ from the tract settings and timbre of control speakers. The differences are explainable in terms of a pharyngeal narrowing that may be congenital or acquired. Other unexplained evidences reported in the literature for specific speech units extracted from OSA patient records may also be explained by this model (e.g. [4], [14], [15]), reinforcing the interpretation offered here.

V. Tracking settings via principal components analysis (PCA)

An alternative way to track vocal tract settings is by means of a principal components analysis. A spectrogram reports the amplitude of the spectral components as a function of frequency on the vertical axis and time on the horizontal axis. With a view to carrying out principal component analysis and inspecting the first principal component [16], the spectrogram is reinterpreted here as a matrix the rows of which are assigned to indexed

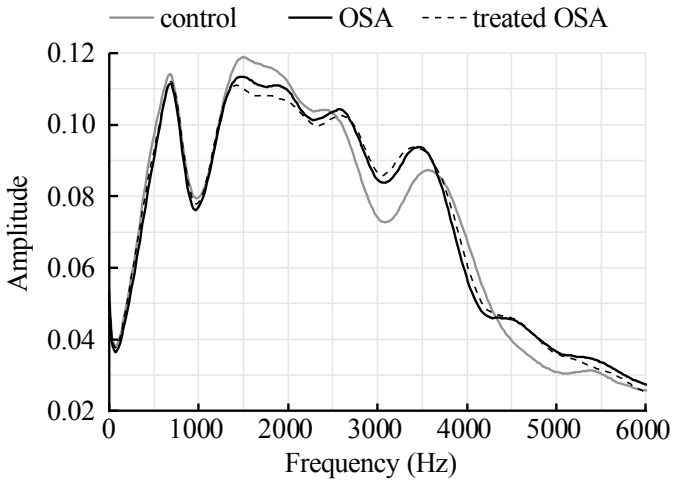

Figure 3: Averaged first principal components for control speakers (continuous, gray line), OSA patients (continuous, black line) and treated OSA patients (dashed, black line) for one single Spanish sentence.

frequency bins and the columns to the index of the analysis frame.

Intuitively speaking, one expects individual amplitude spectra to be a combination of a speaker setting-typical spectrum that is common to all analysis frames, and segment-typical variations that report phonetic identity. When the aggregated segment-typical variations of the spectra are small compared to the setting-typical baseline then the first principal component is expected to report the latter and the higher components the former. The reason is that the mutually uncorrelated principal components are linear combinations of individual amplitude spectra, with the principal components ranked according to the explained variance (i.e. spectral energy).

The feasibility of tract setting analysis via the LTAS suggests that PCA may be suitable for the same task and vice versa. The main difference is that PCA weights analysis frames individually with regard to their phonetic identity, whereas in the LTAS the frame weights are the same.

Fig. 3 shows the averaged first principal components obtained for the control speakers (continuous, gray line), the OSA patients (continuous, black line) and treated OSA patients (dashed, black line). The averaged first principal components are interpretable, similarly to the averages in Figure 2, as spectral contours the F3 and F4 frequencies of which are less distant for OSA speakers than for modal speakers.

ACKNOWLEDGEMENTS

The activities described are partially funded by the Spanish Ministry of Economy and Competitivity as part of the TEC2012-37585-C02 (CMC-V2) Project.

REFERENCES

[1] C. M. Ryan and T. D. Bradley, "Pathogenesis of obstructive sleep apnea," *J. Appl. Physiol.*, vol. 41, no. 6, pp. 323–330, 2005.

[2] F. J. Puertas, G. Pin, J. M. María, and J. Durán, "Documento de consenso Nacional sobre el síndrome de Apneas-hipopneas del sueño," *Grupo Español Sueño*, p. 164, 2005.

[3] J. L. Blanco-Murillo, R. Fernández-Pozo, E. López-Gonzalo, and L. A. Hernández-Gómez, "Exploring differences between phonetic classes in Sleep Apnoea Syndrome Patients using automatic speech processing techniques," *Phon. J. Int. Soc. Phon. Sci.*, vol. 97, pp. 36–55, 2008.

[4] R. Fernández-Pozo, J. L. Blanco-Murillo, L. A. Hernández-Gómez, E. López, J. Alcázar, and D. Torre-Toledano, "Severe Apnoea Detection using Speaker Recognition Techniques.," in *Proceedings of the BIOSIGNALS Conference*, 2009, pp. 124–130.

[5] J. L. Blanco-Murillo, R. Fernández, D. Díaz, L. Hernández, E. López, and D. Torre, "Apnoea Voice Characterization through Vowel Sounds Analysis using Generative Gaussian Mixture Models," in *Proceedings of 3rd Advanced Voice Function Assessment International Workshop*, 2009, vol. 1.

[6] J. Laver, *The phonetic description of voice quality*. Cambridge University Press, 1980.

[7] F. Nolan, *The phonetic bases of speaker recognition*. Cambridge University Press, 1983.

[8] R. Fernández-Pozo, L. A. Hernández-Gómez, E. López-Gonzalo, J. Alcázar-Ramírez, G. Portillo, and D. T.Toledano, "Design of a Multimodal Database for Research on Automatic Detection of Severe Apnoea Cases," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

[9] M. M. Zhao and X. L. Zhang, "Diagnosis and treatment of obstructive sleep apnea hypopnea syndrome," *Zhonghua Yi Xue Za Zhi*, vol. 92, no. 18, pp. 1228–1230, May 2012.

[10] E. Keller, "The Analysis of Voice Quality in Speech Processing," in *Lecture Notes in Computer Science*, 2005, pp. 54–73.

[11] W. J. Hess, "Time-domain digital segmentation of connected natural speech," in *Proceedings of the 4th International Joint Conference on Artificial intelligence*, 1975, vol. 1, pp. 491–498.

[12] I. L. Mortimore, P. Kochhar, and N. J. Douglas, "Effect of chronic continuous positive airway pressure (CPAP) therapy on upper airway size in patients with sleep apnoea/hypopnoea syndrome," *Thorax*, vol. 51, no. 2, pp. 190–192, 1996.

[13] M. Mrayati, R. Carré, and B. Guerin, "Distinctive regions and modes: a new theory of speech production," *Speech Commun.*, vol. 7, no. 3, pp. 257–286, 1988.

[14] J. A. Fiz, J. Morera, J. Abad, A. Belsunces, M. Haro, J. I. Fiz, R. Jane, P. Caminal, and D. Rodenstein, "Acoustic analysis of vowel emission in obstructive sleep apnea," *Chest*, vol. 104, no. 4, pp. 1093–6, 1993.

[15] M. P. Robb, J. Yates, and E. J. Morgan, "Vocal tract resonance characteristics of adults with obstructive sleep apnea," *Acta Otolaryngol*, vol. 117, no. 5, pp. 760–3, 1997.

[16] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philos. Mag.*, vol. 1, no. 11, pp. 559–572, 1901.